
Structural Similarity-based Link Prediction in Complex Networks

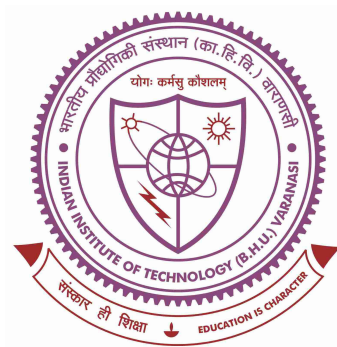
Thesis submitted in partial fulfillment

for the Award of Degree

DOCTOR OF PHILOSOPHY

by

AJAY KUMAR



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

Roll No: 16071005

2021

Certificate

It is certified that the work contained in this thesis entitled “Structural Similarity-based Link Prediction in Complex Networks” by “AJAY KUMAR” has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.

Dr. Bhaskar Biswas

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (Banaras Hindu University),

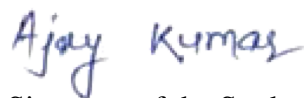
Varanasi-221 005

DECLARATION BY THE CANDIDATE

I, AJAY KUMAR, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Dr. Bhaskar Biswas from July 26, 2016 to October 5, 2021, at the Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Date :

Place : Varanasi



Signature of the Student

(AJAY KUMAR)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Signature of Supervisor

(Dr. Bhaskar Biswas)

Signature of Head of Department

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis : **Structural Similarity-based Link Prediction in Complex Networks**

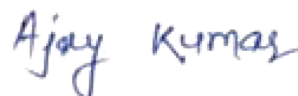
Name of the Student : **AJAY KUMAR**

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the "DOCTOR OF PHILOSOPHY".

Date :

Place : Varanasi



Signature of the Student

(AJAY KUMAR)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

*This thesis is dedicated to my parents and family members
for their endless love, support and encouragement*

PREFACE

Link prediction in complex networks (e.g., social networks, biological networks, citation networks, etc.) has attracted increasing attention from both physical and computer science communities. The algorithms can be used to extract missing information, identify spurious interactions, evaluate network evolving mechanisms, and so on. Its study is crucial to the analysis of the evolution of networks. Lots of works employing different types of methodologies of link prediction are available. Most of them are based on structural or topological properties as extracting these features are easy in computation. Though not all of them are efficient to extract. Most social networks exhibit some basic features like Small-world phenomenon, clustering and scale-free. Their corresponding measures are average path length, clustering coefficient and degree distribution respectively. In this thesis, these features are explored for calculating similarity measures of node-pairs in link prediction.

Many real-world networks show tendency of being organized in clusters that are quantified by clustering coefficient. This measure extracts local structural or topological information which are efficient to compute. The notion of mutual relationships, captured by common neighbors, are building blocks of many existing seminal works like Adamic-Adar index, resource allocation index, etc. The notion of common neighbors is further expanded to higher level. Based on clustering coefficients of level-2 common neighbors, a new algorithm CCLP2 is proposed to predict missing links in networks. CCLP2 extracts higher level clustering information of nodes which proved to be more informative and discriminating feature for link prediction as shown by the empirical results.

Exploring level-2 clustering information are useful discriminating feature but confined to neighbors of neighbors information. This might limit the prediction capability and hence, more local information are extracted using path feature. By employing higher order paths as discriminating features missing link are predicted in networks. The proposed method, called SHOPI, is based on resource allocation process in networks where the source node sends some resources as information to a destination node. The amount of information received by the destination derives the similarity score between

them. Higher the information received by destination from the source represents higher similarity. SHOPI ensures to reach maximum information by restricting the information leaks through their common neighbor nodes. Empirical results on several networks validates the performance of SHOPI.

Acknowledgements

This thesis is the end of my journey in obtaining my Ph.D. I have not traveled in a vacuum in this journey. This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people including my well wishers, my friends, colleagues and various institutions. At the end of my thesis, I would like to thank all those people who made this thesis possible and an unforgettable experience for me. At the end of my thesis, it is a pleasant task to express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me.

Firstly, I would like to express my sincere gratitude to my advisor **Dr. Bhaskar Biswas** for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I specially thank the members of my Research Progress Evaluation Committee, **Dr. R. S. Singh** and **Dr. N. S. Rajput** for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for the hard question which incited me to widen my research from various perspectives.

I would like to express my sincere thanks for **Prof. Rajiv Srivastava**, Head, Computer Science and Engineering Department, for his kindness and valuable support in carrying out the research. I express my sincere thanks to the faculty members, **Prof. K. K. Shukla**, **Prof. A. K. Tripathi**, **Prof. R. B. Mishra**, **Dr. Ravi Shankar Singh**, **Dr. Vinayak Srivastava**, **Dr. R. N. Choudary**, **Dr. H. P. Gupta**, **Dr. T. Dutta**, **Dr. A. K. Singh** and staffs of the department.

I am also thankful to all my friends for the fun moments in my PhD student life. I wish to express my deep sense of thanks to my seniors **Dr. Dharmendra Prasad Mahato**, **Dr. Harish Kumar Shakya**, **Dr. Anupam Biswas**, **Dr. Vinay Kumar**, **Dr. Ashish Kumar Maurya**, and **Dr. Vibhav Singh** for being a great friends and the best advisers I could ever have. Their advice, encouragement and critics were always the source of inspiration. This dissertation would not have been possible without their invaluable suggestions and persistent help. I wish to extend my gratitude towards all my lab mates **Mr. Shashank**

Sheshar Singh, Mr. Shivansh Mishra, Ms. Sneha Mishra, Mr. S. P. Dwivedi, Mr. Naveen Mani Upadhyay, and Mr. Vishal Shrivastva whose company helped me a lot. I am also thankful to **Mr. Yadul Raghav** and **Mr. Aman Gupta** for the collaborations and contributions to some of the projects in this dissertation. I am specially thanks to **Dr. Kuldeep Singh** without him this study might not come to an end.

Most importantly, my deepest gratitude is for my family for their constant support, inspiration, guidance, and sacrifices. My parents were constant source of motivation and inspiration. Their affection and guidance was instrumental in me choosing Engineering and eventually continuing on to my Ph.D. I would like to thank my mother, who was also my childhood teacher and is always there to stands by my side.

Last but not the least, I would like to special thanks is to my wife, elder brother **Mr. Sanjay Kumar** and younger brothers for supporting me unconditionally throughout my study. I sincerely thank all of them who contributed in for helping me to see the light at the end of every scary tunnel during my Ph.D.

Ajay Kumar
- AJAY KUMAR

Contents

Certificate	iii
Declaration by the Candidate	v
Copyright Transfer Certificate	vii
Preface	xi
Acknowledgements	xiii
Contents	xv
List of Figures	xxi
List of Tables	xxv
Abbreviations	xxvii
Symbols	xxix
1 Introduction	1
1.1 Complex networks	1
1.2 Link prediction	2
1.3 Evaluation criteria	4
1.4 Motivation of the thesis	4
1.5 Contribution of the thesis	5
1.5.1 A comprehensive experimental survey of link prediction	5
1.5.2 Influence of higher level clustering features towards link prediction	6
1.5.3 Influence of higher order paths indices towards link prediction . .	7
1.6 Organization of the thesis	7

2	Social Networks Analysis Background	9
2.1	Social networks and graph Theory	9
2.2	General characteristics of social networks	10
2.2.1	Triadic closure and clustering coefficient	10
2.2.2	Small world phenomenon	11
2.2.3	Scale-free networks and preferential attachment	12
2.2.4	Homophily or assortative mixing	13
2.3	Common studied problems of social network analysis	14
2.3.1	Link-based object ranking (LBR)	14
2.3.2	Community detection	14
2.3.3	Finding central nodes	14
2.3.4	Influence maximization	15
2.3.5	Link prediction	15
2.4	Applications of social network analysis	16
	Application specific to link prediction	16
2.5	Computational complexity of social network analysis	17
	Computational Complexity of Similarity-based Link Prediction	18
3	Link Prediction Techniques, Applications, and Performance: A Survey	21
3.1	Similarity-based methods	22
3.1.1	Local similarity indices	23
3.1.1.1	Common neighbors (CN)	23
3.1.1.2	Jaccard coefficient (JC)	24
3.1.1.3	Adamic/Adar index (AA)	24
3.1.1.4	Preferential attachment (PA)	24
3.1.1.5	Resource allocation Index (RA)	25
3.1.1.6	Cosine similarity or Salton index (SI)	26
3.1.1.7	Sorensen index	26
3.1.1.8	CAR-based common neighbor index (CAR)	26
3.1.1.9	Hub promoted index (HPI)	27
3.1.1.10	Hub depressed index (HDI)	28
3.1.1.11	Local naive Bayes-based common neighbors (LNBCN)	28
3.1.1.12	Leicht-Holme-Newman local index (LHNL)	29
3.1.1.13	Node clustering coefficient (CCLP)	29
3.1.1.14	Node and link clustering coefficient (NLC)	30
3.1.2	Global similarity indices	30
3.1.2.1	Katz index	30
3.1.2.2	Random walk with restart (RWR)	31
3.1.2.3	Shortest path	32
3.1.2.4	Leicht-Holme-Newman global index (LHNG)	33
3.1.2.5	Cosine based on L^+ (Cos^+)	34

3.1.2.6	Average commute time (ACT)	34
3.1.2.7	Matrix forest index (MF)	35
3.1.2.8	SimRank (SR)	36
3.1.2.9	Rooted PageRank (RPR)	37
3.1.3	Quasi-local indices	37
3.1.3.1	Local path index (LP)	38
3.1.3.2	Path of length 3 (L3)	38
3.1.3.3	Similarity based on local random walk and superposed random walk (LRW and SRW)	40
	Remarks	41
3.2	Probabilistic and maximum likelihood models	41
3.2.1	Local probabilistic model for link prediction	42
3.2.2	Probabilistic relational model for link prediction (PRM)	44
3.2.3	Hierarchical structure model (HSM) [1]	45
3.2.4	Stochastic block model (SBM) [2]	47
3.2.5	Exponential random graph model (ERGM) or P-star model	50
3.3	Dimension reduction frameworks for link prediction	51
3.3.1	Embedding-based link prediction	52
3.3.2	Factorization-based frameworks for link prediction	55
3.4	Other approaches	57
3.4.1	Learning-based frameworks for link prediction	57
3.4.2	Information theory-based link prediction	58
3.4.3	Clustering-based link prediction	60
3.5	Experimental Setup and Results Analysis	62
3.5.1	Datasets	63
3.5.2	Accuracy	64
	Recall@k	65
	Area under the precision-recall curve (AUPR)	65
	Area under the receiver operating characteristics curve (AUROC)	66
	Average precision	68
	Concluding remarks	69
	Parameters settings	70
3.5.3	Efficiency	71
3.6	Variations of link prediction problem	73
3.6.1	Link prediction in weighted and directed networks	73
3.6.2	Link prediction in temporal networks	75
3.6.3	Link prediction in bipartite networks	76
3.6.4	Link prediction in heterogeneous networks	77
3.7	Link prediction applications	78
3.7.1	Network reconstruction	78

3.7.2	Recommender system	80
3.7.3	Network completion problem	81
3.7.4	Spam mail detection	82
3.7.5	Privacy control in social networks	82
3.7.6	Identifying missing references in a publication	83
3.7.7	Routing in networks	83
3.7.8	Incorporating user's influence in link prediction	84
4	Level-2 Node Clustering Coefficient-based Link Prediction	87
4.1	Introduction	87
4.2	Proposed work	89
	Extracting more local information	90
	Link prediction based on Level-2 node clustering coefficient.	92
	Algorithm description.	94
4.3	Experimental study	94
4.3.1	Evaluation metrics	94
4.3.2	Datasets description	96
4.3.3	Results analysis	98
	AUROC	98
	AUPR	99
	AP	102
	Recall	102
	Concluding remarks.	102
	Complexity analysis.	104
	Statistical test	106
4.4	Conclusion and future works	107
5	Link Prediction in Complex Networks Based on Significance of Higher-Order Path Index (SHOPI)	111
5.1	Introduction	111
5.2	Proposed work	115
	Significance of the path index of length 2	116
	Significance of path index of higher order	116
	Algorithm description	118
5.3	Experimental study	119
5.3.1	Evaluation metrics	119
5.3.2	Datasets description	122
5.3.3	Results analysis	123
	AUROC	123
	Average precision (AP)	125
	Effects of the parameter value ψ and sensitivity analysis	126
	AUROC sensitivity	126

AP sensitivity	126
Significance of higher order paths	129
Complexity analysis.	129
Statistical test	131
5.4 Conclusion	131
6 Conclusion and future directions	135
6.1 Conclusion	135
6.2 Future directions	136
Bibliography	139
A List of Publications	173

List of Figures

1.1	The Link Prediction (LP) finds missing links (i.e., AD, AC, and BD) in this observed network.	3
2.1	Triads in networks	10
2.2	Small world phenomenon with pure essential growth	12
3.1	Taxonomy of Link Prediction Approaches	22
3.2	$CAR\ Index = (Number\ of\ CNs) \times (Number\ of\ LCLs)$	27
3.3	Local probabilistic model for link prediction [3]	42
3.4	An illustrating example of HSM for a graph of 6 nodes and its two possible dendrograms as described in the paper [1]. The internal nodes of each dendrogram are labeled as the maximum likelihood probability \bar{p}_r , defined by the equation 3.45. The likelihoods of the left and the right dendrograms are $L(D_1) = (1/3)(2/3)^2 \cdot (1/4)^2(3/4)^6 = 0.00165$, and $L(D_2) = (1/9)(8/9)^8 = 0.0433$. Thus, the second (i.e., right) dendrogram is most probable as it divides the network in a balanced one at the first level.	48
3.5	The Karate club network (left) and its representation in the embedding space with the DeepWalk [4] algorithm.	52
3.6	Embedding of nodes x and y to the embedding space	53

3.7	An example illustrating the cycle formation link probability model [5], where the the probability of the missing link $(x - y)$ is generated by the following three mechanisms; random link occurrence $g(1)$, length-2 cycle generation $g(2)$ i.e. $(x - a - y, x - c - y)$, and length-4 cycle generation $g(3)$ i.e. $(x - b - d - y)$	61
4.1	Initially at time t_0 , three links are present in the disconnected graph. As the time progress, more links are formed as shown at the time instant $t_{(k-1)}$. Now, at the time instant t_k , which of the non-existing links (i.e. AC, AE, BC, BE, CD) will be formed ? Finding the potential links that will appear at t_k is called the link prediction problem.	88
4.2	Notion of level-2 clustering coefficient	90
4.3	Exploring local to global structure	91
4.4	Computing level-2 node clustering coefficient	92
4.5	AUROC Results	100
4.6	AUPR Results	101
4.7	AP Results	103
4.8	Recall Results	105
5.1	Path-based approaches to link prediction	113
5.2	Path length-2 score calculation: the score between x and y , $S(x, y) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$, and the score between p and q , $S(p, q) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	116
5.3	Three possible paths (blue colored edges) between the node pair (x, y) during the computation of path length-3 score.	118
5.4	Confusion Matrix	120
5.5	AUROC results sensitivity corresponding to different parameter values of ψ	127

5.6 AP results sensitivity corresponding to different parameter values of ψ . . 128

List of Tables

3.1	Comparison of similarity-based approaches	41
3.2	Probabilistic and maximum likelihood models for link prediction	46
3.3	Deep learning models for embedding based link prediction	54
3.4	Topological information of real-world network datasets	64
3.5	Recall Results	66
3.6	AUPR Results	67
3.7	AUROC Results	68
3.8	Average Precision Results	69
3.9	Recall results for other representative methods	70
3.10	AUPR results for other representative methods	70
3.11	AUROC results for other representative methods	71
3.12	Average precision results for other representative methods	71
3.13	The computational Complexity of similarity-based methods and the corresponding references	72
3.14	Link Prediction in Temporal Networks	76

3.15	Link Prediction in Heterogeneous Networks	79
4.1	Topological information of real-world network datasets	97
4.2	The Friedman test on Area under the ROC Curve (AUROC)	108
4.3	The Posthoc Friedman Conover Test (Control method = CCLP2)	109
5.1	Topological information of real-world network datasets	123
5.2	AUROC Results	125
5.3	Average Precision (AP) Results	125
5.4	Effects of considering longer path lengths on the accuracy of link prediction	129
5.5	The Friedman test on area under the ROC Curve (AUROC) and average precision (AP)	132
5.6	The Posthoc Friedman Conover Test (Control method = SHOPI, Correction method = Holm)	132

Abbreviations

CCLP2	Level-2 Node Clustering Coefficient-based Link Prediction
SHOPI	Link Prediction in Complex Networks based on Significance of Higher-Order Path Index
AUROC	Area Under the Receiver Operating Characteristic Curve
AP	Average Preccision
AUPR	Area Under the Precision-Recall Curve
LCP	Local Community Paradigm
LCL	Local Community Links
CAR	Cannistraci-Alanis-Ravasi
L3	Path of length 3
NSI	Neighbor Set Index
CCLP	Clustering Coefficient-based Link Prediction
NMF	Non-negative Matrix Factorization
DCP	Degree related Clustering ability Path
DR	Degree of Robustness

Symbols

$G(V, E)$	A social network with vertex set V and edge set E
A	Adjacency matrix of a network
n	The number of nodes in the network ($ V $)
m	The number of Edges in the network ($ E $)
$\Gamma(z)$	The neighbors set of node z
k_z	Degree of the node z
$t(z)$	Number of triangle passing through the node z
λ_1	Maximum eigen value of a matrix
β	damping factor
$\langle K \rangle$	Average degree of the a network
$\langle D \rangle$	Average path length of a network
$\langle C \rangle$	Average clustering coefficient of a network
r	Coefficient of assortativity of a network
H	Degree of heterogeneity of a network
ρ	Network density
D	Diagonal matrix
\mathcal{U} and \mathcal{V}	are left and right singular vectors
L	Laplacian matrix
$C(z)$	Clustering coefficient of node z
CN^2	Level-2 common neighbors
$CC(z)$	Level-2 clustering coefficient a node z

α_c	Level of confidence
D_f	Degree of freedom
$S(x,y)$	Similarity score between the node x and the node y
ψ	Penalization factor for longer paths
l_{max}	Length of maximum path in the network
i_1, i_2	Intermediate nodes
$I_1 \dots I_8$	Information flow in the network
I	Identity matrix